WRF benchmark for regional applications

[1,2]D. Arnold, [3]D. Morton, [1]I. Schicker, [4]O. Jorba, [3]K. Harrison, [5]J. Zabloudil, [3]G. Newby, [1]P. Seibert

1 Institute of Meteorology, University of Natural Resources and Life Sciences, Vienna, Austria
2 Institute of Energy Technologies, Technical University of Catalonia, Barcelona, Spain
3 Arctic Region Supercomputing Center, University of Alaska Fairbanks, USA
4 Barcelona Supercomputing Center, Barcelona, Spain
5 Vienna Supercomputing Center, Vienna, Austria

ABSTRACT: *Despite parallelisation, scalable performance of WRF is often not achieved. Although the WRFV3 Parallel Benchmark Page provides valuable scaling information, this one-domain configuration is often not typical of the general application of WRF to nested domains of varying sizes and shapes. This study is a first step in providing a centralised WRF performance repository for a variety of typical configurations and environments that might be found in general NWP environments. Here, a real-world scenario that comes close to the ones used in regional climate studies in complex topographical areas is evaluated. The performance of WRF V3.2.1 is analysed in several multi-nest configurations - varying vertical levels and parameterisations - and tested for scalability on different platforms. Initial insight shows that multi-nest configurations make realization of desired scalabilities more difficult. In many real-world cases, the outer nest has substantially fewer grid points than the inner nests, and this can be problematic in that large-scale domain decomposition of the inner nests - where the great majority of computations take place - is limited by "over-decomposition" of the outer nest.*

KEYWORDS: WRF, benchmarking

1. Introduction

The Weather Research and Forecasting model (Skamarok et al., 2008) is extensively used by the modelling community on a wide range of applications and degrees of complexity. In such applications, there is typically a preliminary step in which the platform to use, the resources needed and the feasibility of the study are evaluated. To aid the modellers to achieve this aim, comprehensive benchmark studies, data-sets and evaluation tools should be accessible. The WRFV3 Parallel Benchmark page (http://www.mmm.ucar.edu/wrf/WG2/bench/) already provides the necessary information to the public to carry out a performance analysis of WRF version 3.0 for two different single-domain configurations. However, this configuration may not be representative of the more complicated and demanding multi-nest multi-shaped configurations regional applications may need, which could easily consist of 3 to 5 nesting levels going down to resolutions of 1 km. Some recent work has also dealt with the evaluation of WRF performance when a more complex configuration is defined (Porter et al, 2010) and has encouraged the starting of this study. In this paper, a real-world scenario, close to the ones used at the Institute of Meteorology at the University of Natural Resources and Life Sciences for regional climate studies, is implemented and tested in different platforms and configurations. An area of initial concern was the potential load imbalances that might occur. In a typical nested WRF configuration, outer nests tend to have much fewer grid points than the inner nests and, because WRF will use the same number of parallel tasks to integrate each nest, there was concern that the fine-grained parallel domain decomposition on the outer nests might take away from performance gained by deploying a large number of tasks on the denser inner nests.

2. Benchmark configuration

2.1 Domain set-up

Three different nest configurations over Europe were initially considered. The first one, henceforth called *4dbasic* case (Table 1, Figure 1), consists on four 1-way nested domains with the innermost one covering the northern part of the Alps and each of them with 40 vertical levels. The second case, *4dbasiclev*, is basically as the *4dbasic* configuration but with an increase in the vertical levels from 40 to 63 keeping the same relative horizontal domain position and size. This results in over 50% more grid points in each nest. The last configuration, *3dhrlev* (Table 2, Figure 2), is a three-domain set-up with the outer domain of *4basiclev* at 7.2 km horizontal resolution, nesting down to 2.4 km and 800 m resolution, discarding domain 4. The large innermost domain covers now all the Alpine region.

Domain	Grid cells	Total cells	Horizontal resolution
1	196x167x40	1.3 M	21.6 km
2	274x217x40	2.4 M	7.2 km
3	274x217x40	8.4 M	2.4 km
4	1003x505x40	20.3 M	800 m

Table 1: Description of the *4dbasic* configuration.

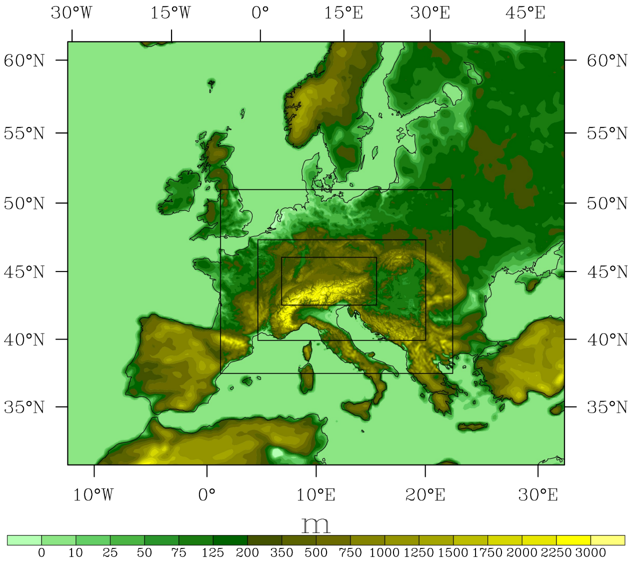

Figure 1: 4-domain configuration for the *4dbasic* and *4dbasiclev* cases.

These configurations present several challenges to the model. Amongst them, two need to be highlighted here: the computational costs due to the large amount of grid cells, and stability of the model. The last one proved to be crucial already in the initial steps of the study since the the steep slopes of the inner domains often lead to unrealistic vertical velocities.

Domain	Grid cells	Total cells	Horizontal resolution
1	585x495x63	16.9 M	7.2 km
2	823x652x63	33.8 M	2.4 km
3	1777x1066x63	119.3 M	800 m

Table 2: Description of the *3dhrlev* configuration.

2.2 Computational platforms

The four computational platforms used for the initial testing are briefly outlined below, describing only the resources used in this project (some of the systems are heterogeneous):

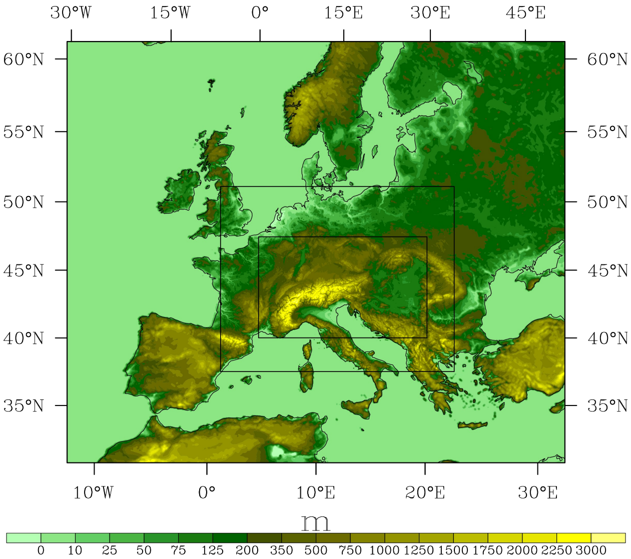

Figure 2: 3-domain configuration for the *3dhrlev* case.

- **Vienna Scientific Cluster**: operated to satisfy the demand for High Performance Computing at the University of Vienna, the Vienna University of Technology, and the University of Natural Resources and Life Sciences.
  - Sun Fire X2270 compute nodes, each equipped with 2 Quadcore processors (Intel, X5550, 2.66 GHz) and 24 GB memory (3 GB per core).
  - Infiniband QDR network (40 Gbps).
  - Filesystem – ext2.

- **Pacman**: academic system at the Arctic Region Supercomputing Center, funded by NSF in support of the Pacific Area Climate Monitoring and Analysis Network (PACMAN). Penguin Computing Cluster with:
  - Sixteen-core compute nodes consisting of 2 eight-core 2.3 GHz AMD Opteron processors with 64 GB memory (4 GB per core).
  - Mellanox QDR Infiniband interconnect.
  - Panasas version 12 file scratch file system.

- **Kraken**: Cray XT5 at National Institute for Computational Sciences.
  - 12-core compute nodes consisting of 2 six-core 2.6 GHz AMD Opteron processors with 16 GB memory (1.5 GB per core).
  - Cray SeaStar2+ interconnect.
  - Lustre filesystem used on compute nodes.

- **Chugach**: Cray XE6 currently administered by ARSC for the DoD High Performance Computing and Modernization Program.

- 16-core compute nodes consisting of 2 eight-core 2.3 GHz AMD Opteron processors with 32 GB memory (2 GB per core).
- Cray Gemini interconnect.
- Lustre scalable filesystem used on compute nodes.

2.3 Performance assessment

Several metrics has been sued to asses the performance of the runs:

- **Total wall-time**: this is simply the time from job initiation to termination.
- **Integration wall-time**: the WRF program produces an auxiliary output file which annotates the time required for each timestep, for each nest, and includes I/O times. A Python script has been created that accumulates the integration times for Nest 1 (which naturally include integration times for underlying nests) and therefore reflect a rough estimate of "time spent computing."
- **I/O wall time**: computed as the difference between total wall-time and integration wall-time as an approximate proxy.

It needs to be highlighted that this initial performance evaluation was studied on single runs and not on averaged multiple runs, leading to possible noise in the data. In addition, compilers and compilation options were not the same for all the platforms, making the strict comparison amongst them impossible.

From this information the scalability and speed-up for each of the runs and machines can be easily computed scaled with the smallest (lower core number) test case, since in most cases a serial case would be too memory intensive for single-task execution.

3. Results and discussion

For each of the platforms the scalability and speed-up have been calculated for all the cases. For the VSC runs (Figures 4 and 5), as for the rest of the platforms, a similar behaviour can be observed, whereby total wall time, which includes I/O, tends to flatten out a bit more rapidly than the integration wall time. This indicates that for large problem sizes as the ones dealt with, I/O operations sometimes take longer than the actual computations. It can be observed a superlinear speedup for the *3dhrlev* case, which can be attributed to irregular performance of the VSCr under varying load conditions, recognizing that such variation is normal in many computing clusters.

The relatively new and uncrowded Chugach, a Cray XE6, presents fairly stable performance (Figures 6 and 7), in agreement with previous studies performed by the Arctic Region Computer Center (ARSC) on many other Cray machines (Morton et al., 2009). Note that the Chugach cases used up to 2048 tasks, and there appears to be a steady plateauing of the total wall times, reflecting the increasing I/O bottlenecks encountered as we have more tasks trying to coordinate the writing of a single output file. In all of these cases, master/slave I/O has been employed, and some preliminary results suggest that I/O performance improves markedly with some parallel I/O approaches (Porter et al., 2010, Li et al., 2003).

Performance on Pacman and Kraken (Figures 8-11) tends to be more "noisy," and with a dominance of the I/O costs. Of particular interest is the large and demanding *3dhrlev* case on Kraken (note that the upper scale of Figure 11 is used, reflecting job runs of 2400 to 5760 processing elements). The I/O costs are very high in this case, yet when removed (by considering integration wall time) there is still reasonable scalability. This encourages to further test and use I/O schemes other than the basic default master/slave paradigm used for these tests.

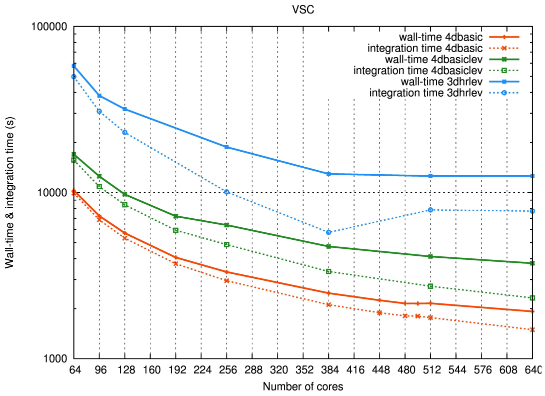

Figure 4: VSC total wall and integration time versus cores

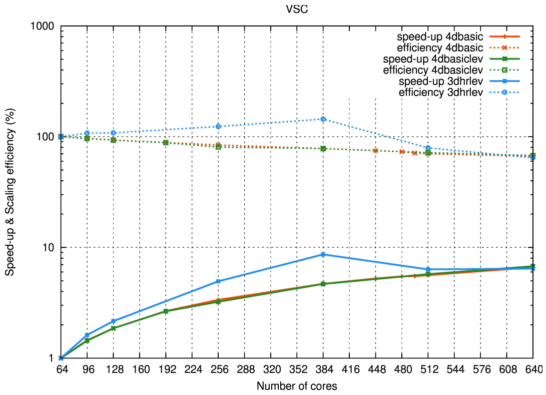

Figure 5: VSC speedup and efficiency versus cores

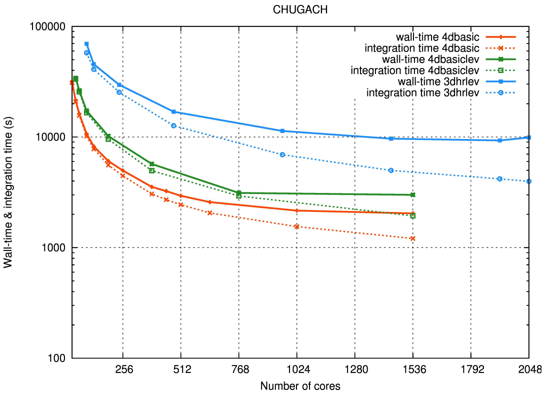

Figure 6: Chugach total wall and integration time versus cores

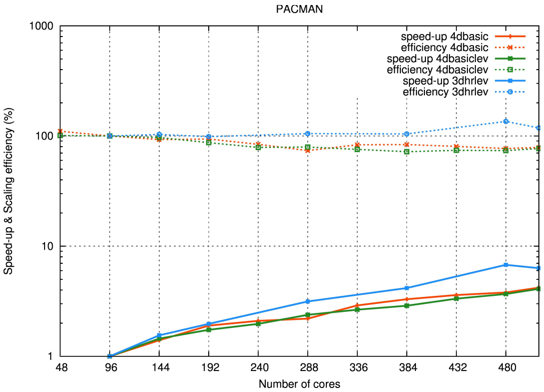

Figure 9: Pacman speedup and efficiency

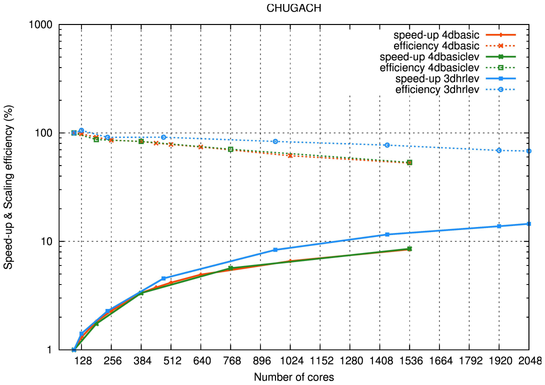

Figure 7: Chugach speedup and efficiency

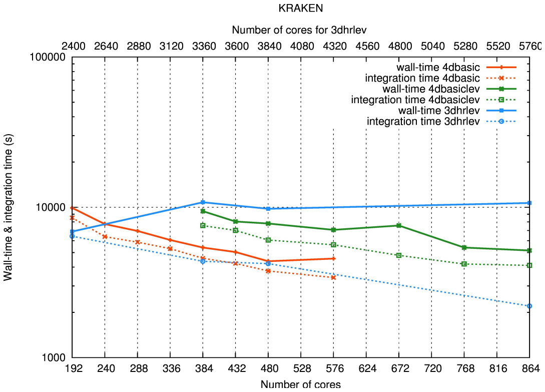

Figure 10: Kraken total wall and integration time versus cores

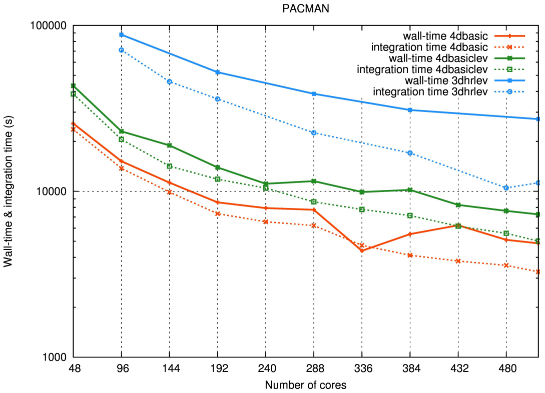

Figure 8: Pacman total wall and integration time versus cores

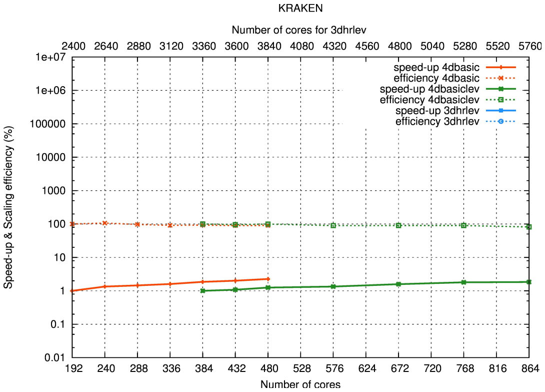

Figure 11: Kraken speedup and efficiency

One of the outcomes of more interest for those aiming at these sort of WRF configurations, with a nested configuration using a rather coarse outer domain with much fewer grid points than its inner nests, is that the scalability does not seem to suffer. Given that under typical 3:1 nesting scenarios, child nests will incur three times the number of time-steps than their parent nest, and they will tend to possess a larger number of grid points. Though more rigorous analysis is warranted, it appears that the time required to integrate a particular nest is somewhat negligible when compared to the time required for its child nest and the same for the next nesting level. A preliminary inspection of the simulations of this study, reveal that most of the integration time is actually spent in the inner, more demanding, nest, decreasing the importance of possible inefficiencies that might result from fine-scale domain decomposition of the parent nests. A significant pitfall that the user needs to be aware of, however – and this was encountered in some the runs with large numbers of tasks - is that the same large number of tasks applied to a higher level nest may result in over-decomposition, whereby individual tasks simply do not have enough grid points to work with (with halo points for communications, WRF requires each task to have some certain number of grid points), and then the process crashes.

5. Towards a versatile benchmark suite

As seen elsewhere in the literature, different WRF benchmark studies have been presented for some very specific cases, generally with one single domain configuration. However, a large constellation of configurations are defined by the community according to the user needs, region of interest and final application. With this in mind, the ARSC and the Institute of Meteorology from the University of Natural Resources and Life Sciences of Vienna, have joined efforts to create a WRF benchmark page addressing two of the issues that challenge the computational performance of WRF, namely, the use of very large domains with large number of grid points, and the simulations of nested configurations over places with complex topography that would require, in principle, some nesting levels to achieve the desired horizontal resolution. This will be based on the already existing benchmark page at ARSC (http://weather.arsc.edu/BenchmarkSuite/) in which single nest domains with increasing resolution are provided to the user for their testing.

In this new more versatile benchmark page the user would find:

- Data sets for different degrees of complexity and as versatile as possible within, of course, reasonable constrains to keep data sets at a reasonable size.
- Evaluation tools and scripts to easily get results and improve comparability.
- A platform to communicate and share experiences since domains and set-ups might be solicited or posted by the users.

Acknowledgments

Part of this work has been carried out as a part of the Austrian Climate Research Programme (ARCP) part of the Klima und Energiefonds (Kli.En). The computational results presented have been achieved using the Vienna Scientific Cluster (VSC).

This work was supported in part by a grant of HPC resources from the Arctic Region Supercomputing Center at the University of Alaska Fairbanks as part of the Department of Defense High Performance Computing Modernization Program.

This research was supported by an allocation of advanced computing resources provided by the National Science Foundation. The computations were performed on Kraken at the National Institute for Computational Sciences (http://www.nics.tennessee.edu/)

References

Jianwei L., W. Liao, A. Choudhary, R. Ross, R. Thakur, W. Gropp, R. Latham, A.w Siegel, B. Gallagher, and M. Zingale. Parallel netCDF: A Scientific High-Performance I/O Interface. *Proceedings of the 15th Supercomputing Conference*, Phoenix, AZ, November 2003.

Michalakes, J. et al, "WRF nature run," *Journal of Physics: Conference Series*, Volume 125, Number 1, SciDAC 2008, 13-17 July 2008, Washington, USA,2008.

Morton, D., O. Nudson, and C. Stephenson, "Benchmarking and Evaluation of the Weather Research and Forecasting (WRF) Model on the Cray XT5" in *Cray User Group Proceedings*, Atlanta, GA, 04-07 May 2009

Morton, D.,O. Nudson, D. Bahls and G. Newby, "Use of the Cray XT5 Architecture to Push the Limits of WRF Beyond One Billion Grid Points" in *Cray User Group Proceedings*, Edinburgh, Scotland, 24-27 May 2010.

Porter A.R., M. Ashworth, A. Gadian, RT. Burton, P. Connolly, M. Bane. WRF code Optimisation for Mesoscale Process Studies (WOMPS) dCSE Project Report, June 2010.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, M. Duda, X.-Y. Huang, W. Wang and J. G. Powers, "A Description of the Advanced Research WRF Version 3", NCAR Technical Note, 2008.